Statistica descrittiva

Laboratorio di Bioinformatica
Corso A
aa 2005-2006

La Statistica si occupa dell'analisi quantitativa dei fenomeni collettivi (cioè fenomeni composti da un grande numero di unità elementari).

Esempi di fenomeni collettivi:

- **L'insieme degli studenti di un corso universitario.**
  Quali sono le loro caratteristiche?
- **L'insieme dei potenziali pazienti che soffrono di ipertensione.**
  Il farmaco A è più efficace del farmaco B?

Gli scopi della statistica sono quindi

- Descrivere
- Generalizzare
- Prevedere

La statistica è l'insieme dei metodi, fondati sul calcolo delle probabilità, che consentono, da un lato la corretta programmazione di un esperimento o di una osservazione pianificata e, dall'altro, l'elaborazione dei dati così raccolti.

La statistica moderna può essere divisa in tre parti:

- Statistica descrittiva
- Statistica matematica
- Statistica inferenziale

La Statistica descrittiva

- Lo scopo della statistica descrittiva è quello di descrivere efficacemente una grande massa di dati mediante tabelle e grafici e di sintetizzare le informazioni in indici matematici in modo da individuare le caratteristiche fondamentali del campione

La Statistica matematica

- La Statistica matematica si avvale del Calcolo delle Probabilità e presenta le distribuzioni teoriche per misure discrete e continue

La Statistica inferenziale

- La Statistica inferenziale si occupa di dedurre leggi generali disponendo di un campione variabile. In pratica è l'insieme dei metodi che consentono di pervenire a delle conclusioni che vanno al di là della stretta evidenza empirica

Il linguaggio della Statistica descrittiva

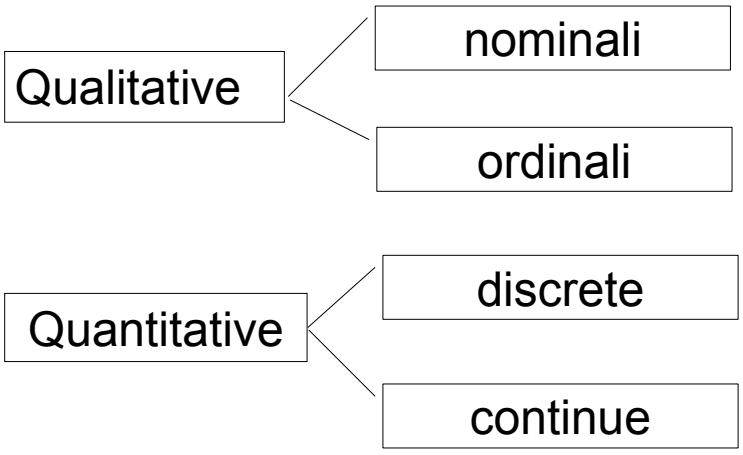
- **Popolazione statistica**: è l'insieme di tutti i possibili oggetti dell'indagine statistica
- **Individuo** (o **unità statistica**): è un qualsiasi elemento della popolazione
- **Variabile**: è una qualsiasi caratteristica di ogni individuo della popolazione, soggetta a variazioni di valore da un individuo all'altro

Indagine sulle domande di adozione nel distretto della Corte d'Appello di Torino nel 2003 (Istat)

- Tipo di indagine: censimento
- Popolazione: coppie che hanno presentato domanda di adozione nel distretto della Corte d'Appello di Torino nel 2003
- Individuo: una qualunque coppia

- Variabili: domande poste alle coppie mediante un questionario:
  - Età dei coniugi
  - Titolo di studio dei coniugi
  - Reddito familiare
  - Tipo di matrimonio
  - Numero di figli
  - Tipo di adozione (nazionale o internazionale)

Classificazione delle variabili



Età dei coniugi	quantitativa discreta (?)
Titolo di studio dei coniugi	qualitativa ordinale
Reddito familiare	quantitativa continua
Tipo di matrimonio	qualitativa nominale
Numero di figli	quantitativa discreta
Tipo di adozione	qualitativa nominale

Coppie che hanno presentato domanda di adozione alla corte di appello di Torino – anno 2003

n. coppia	1	2	3	4	5	6
età marito	35	42	38	51	32	…
età moglie	34	36	39	45	30	…
Tit.studio marito	LAUREA	DIP. SUP.	DOTTO RATO	LIC. MEDIA	DIP. SUP.	…
Tit. studio moglie	LAUREA	DIP. SUP.	LAUREA	DIP. SUP.	LAUREA	…
Tipo di matrimonio	RELIG.	RELIG.	CIVILE	RELIG.	RELIG.	…
Reddito	40.700	35.850	45.225	35.000	30.315	…
Numero Figli	0	1	0	1	0	…
Tipo Adozione	INTERN.	INTERN.	NAZ.	INTERN.	INTERN. E NAZ.	…

FREQUENZA

- La frequenza di un valore è il numero di individui della popolazione per i quali la variabile assume tale valore

TITOLO DI STUDIO DELLA MOGLIE

TITOLO DI STUDIO	FREQUENZA
Dottorato o specializ.	15
Laurea	139
Diploma universitario o laurea breve	22
Diploma di scuola media superiore	249
Licenza di scuola media inferiore	113
Licenza elementare	3
Non indicato	4
Totale	545

TITOLO DI STUDIO DELLA MOGLIE

TORINO

TITOLO DI STUDIO	FREQUENZA
Dottorato o specializ.	15
Laurea	139
Diploma universitario o laurea breve	22
Diploma di scuola media superiore	249
Licenza di scuola media inferiore	113
Licenza elementare	3
Non indicato	4
Totale	545

FIRENZE

TITOLO DI STUDIO	FREQUENZA
Dottorato o specializ.	16
Laurea	65
Diploma universitario o laurea breve	18
Diploma di scuola media superiore	160
Licenza di scuola media inferiore	72
Licenza elementare	4
Non indicato	2
Totale	337

FREQUENZA RELATIVA

- La frequenza relativa è il rapporto tra la frequenza del valore e il numero di individui della popolazione:
  freq. relat. = freq. ass. / totale individui
- La frequenza percentuale si ottiene normalizzando a 100 il totale della popolazione:
  freq. percentuale = freq. relativa * 100

FREQUENZE RELATIVE

TORINO

TITOLO DI STUDIO	FREQUENZA RELATIVA	FREQUENZA PERCENTUALE
Dott. o spec.	0,0275	2,75%
Laurea	0,2550	25,50%
Diploma univers.	0,0404	4,04%
Diploma superiore	0,4569	45,69%
Licenza media	0,2073	20,73%
Licenza elem.	0,0055	0,55%
Non indicato	0,0073	0,73%
Totale	1	100%

FIRENZE

TITOLO DI STUDIO	FREQUENZA RELATIVA	FREQUENZA PERCENTUALE
Dott. o spec.	0,0475	4,75%
Laurea	0,1929	19,29%
Diploma univers.	0,0534	5,34%
Diploma superiore	0,4748	47,48%
Licenza media	0,2136	21,36%
Licenza elem.	0,0119	1,19%
Non indicato	0,0059	0,59%
Totale	1	100%

FREQUENZE CUMULATIVE (TORINO)

TITOLO DI STUDIO	FREQ	FREQ. RELAT.	FREQ. PERC.	FREQ. CUMUL.	FREQ. CUM. %
Dott. o spec.	15	0,0275	2,75%	0,0275	2,75%
Laurea	139	0,2550	25,50%	0,2825	28,25%
Diploma univers.	22	0,0404	4,04%	0,3229	32,29%
Diploma superiore	249	0,4569	45,69%	0,7798	77,98%
Licenza media	113	0,2073	20,73%	0,9871	98,71%
Licenza elem.	3	0,0055	0,55%	0,9926	99,26%
Non indicato	4	0,0073	0,73%	1	100%
Totale	545	1	100%		

Distribuzione

La funzione che ad ogni valore della variabile associa la sua frequenza (o frequenza relativa) si dice distribuzione della variabile.

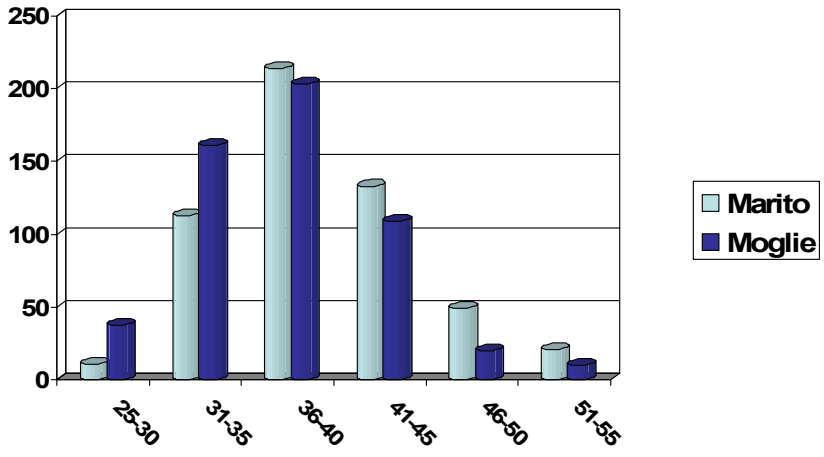
Attenzione: se la variabile e' continua o se i possibili valori sono troppi, si dividono in classi

Esempio: età del marito

CLASSE	FREQ.	FREQ. REL.	FREQ. CUMUL.	
Da 26 a 30	11	2,02%	2,02%	< 30
Da 31 a 35	113	20,73%	22,75%	< 35
Da 36 a 40	214	39,27%	62,02%	<40
Da 41 a 45	133	24,40%	86,42%	<45
Da 46 a 50	49	8,99%	95,41%	<50
Da 51 a 55	21	3,85%	99,27%	<55
Non indicato	4	0,73%	100,00%	
Totale	545	100,00%		

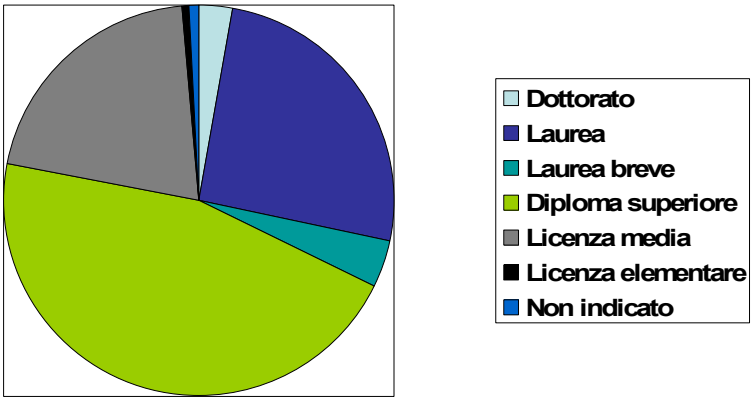
Rappresentazioni grafiche

Istogramma



Rappresentazioni grafiche

Diagramma a torta



INDICATORI SINTETICI

MISURE DI TENDENZA CENTRALE

Sono quantità che individuano i valori intorno ai quali i dati sono raggruppati.

- MEDIA
- MODA
- MEDIANA

Media Aritmetica Semplice

Esempio: “Rossi ha la media del 25”

Popolazione: insieme degli esami sostenuti da Rossi

Variabile: voto ottenuto nell’esame

Media aritmetica semplice =

somma dei voti ottenuti / numero esami sostenuti

Media aritmetica semplice

N = numero di individui di una popolazione

X = variabile numerica

x_i = valore che la variabile assume sull'i-esimo individuo della popolazione

La media è definita da

$$\overline{X} = \frac{1}{N}\sum_{i=1}^{N} x_i = \frac{(x_1 + x_2 + ... + x_N)}{N}$$

La media aritmetica può essere calcolata anche conoscendo solo la distribuzione della variabile.

Siano x_j, per j=1,…, m, i valori che la variabile X può assumere e siano f_j le corrispondenti frequenze. Allora

$$\overline{X} = \frac{\sum_{j=1}^{m} x_j \cdot f_j}{\sum_{j=1}^{m} f_j}$$

Voti ottenuti negli esami

25 27 23 25 23 27 25

M. aritm. =(25+27+23+25+23+27+25)/7=25

(23*2+25*3+27*2)/(2+3+2)=25

Quando la variabile è suddivisa in classi, ad ogni classe si associa il valore medio dell'intervallo

CLASSE	FREQ.	VALORE MEDIO
Da 26 a 30	11	28
Da 31 a 35	113	33
Da 36 a 40	214	38
Da 41 a 45	133	43
Da 46 a 50	49	48
Da 51 a 55	21	53
Totale	541	

$$\overline{X} = \frac{28 \cdot 11 + 33 \cdot 113 + 38 \cdot 214 + 43 \cdot 133 + 48 \cdot 49 + 53 \cdot 21}{541} = 39{,}4$$

Media armonica

$$\frac{1}{H} = \frac{1}{N}\left(\frac{1}{x_1} + \frac{1}{x_2} + \ldots + \frac{1}{x_N}\right)$$

Questa media è la stima più corretta per distribuzioni di dati dei quali devono essere usati gli inversi

La città A dista 100 km dalla città B; andiamo da A a B con un'auto che viaggia a 50 km/h e torniamo con una che viaggia a 70 km/h.

Quanto tempo impieghiamo?

T = 100/50 + 100/70 = 3.43 h

Media aritmetica delle velocità=60 km/h

t = 2*100/60 = 3.33 h

Media armonica=$2(1/50+1/70)^{-1}$=58,33 km/h

t = 2*100/58,33 = 3.43 h

Media geometrica

$$G = \left(x_1 \cdot x_2 \cdot \ldots \cdot x_N \right)^{1/N}$$

Questa media è adatta, per esempio a stimare i tassi di interesse o di inflazione.

Supponiamo che l'inflazione annua in tre anni successivi sia stata del 2,5%, 2% e 1,5%.

Quanto costa un bene che aveva prezzo p?

p'=p*(1.025)*(1.02)*(1.015)=p*1.0611825

Media aritmetica=1.02

p'=p*(1.02)*(1.02)*(1.02)=p*1.061208

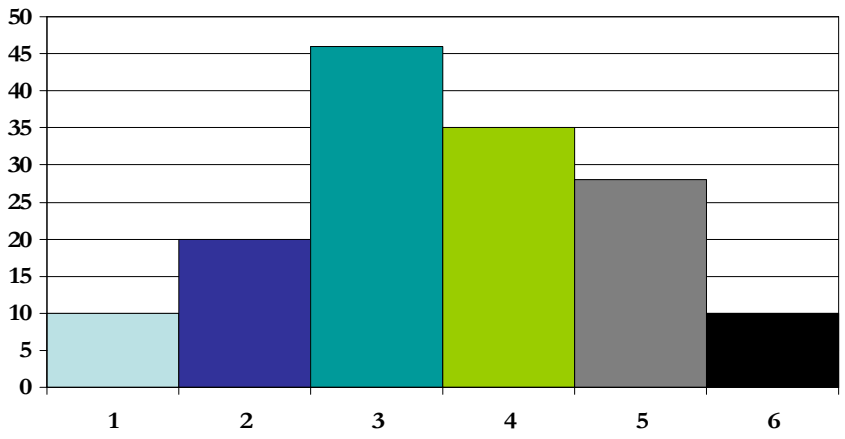
Media geom.=$(1.025*1,02*1,015)^{1/3}$=1.0199

p'=p*(1.0199)*(1.0199) *(1.0199)=p*1.0611825

Moda

- La moda è il valore più frequente di una distribuzione. Può essere definita anche per variabili qualitative.
- Una distribuzione può avere due (o più) massimi di frequenze paragonabili. Si parla allora di distribuzione bimodale.

Moda



Mediana

- La mediana è il valore che occupa la posizione centrale in un insieme ordinato di dati. E' definita solo per variabili ordinali.
- In una distribuzione o serie di dati, ogni valore estratto a caso ha la stessa probabilità di essere inferiore o superiore alla mediana.

Come si calcola la mediana

- Si dispongono i dati in ordine crescente o decrescente e se ne conta il numero totale **n**
- Se **n** è dispari la mediana corrisponde al valore che occupa la posizione centrale **(n+1)/2**
- Se **n** è pari la mediana è la media tra i valori nelle posizioni **n/2** e **(n+2)/2**

Confronto media e mediana

Serie:

23 45 67 73 96 108 132 156 177

mediana

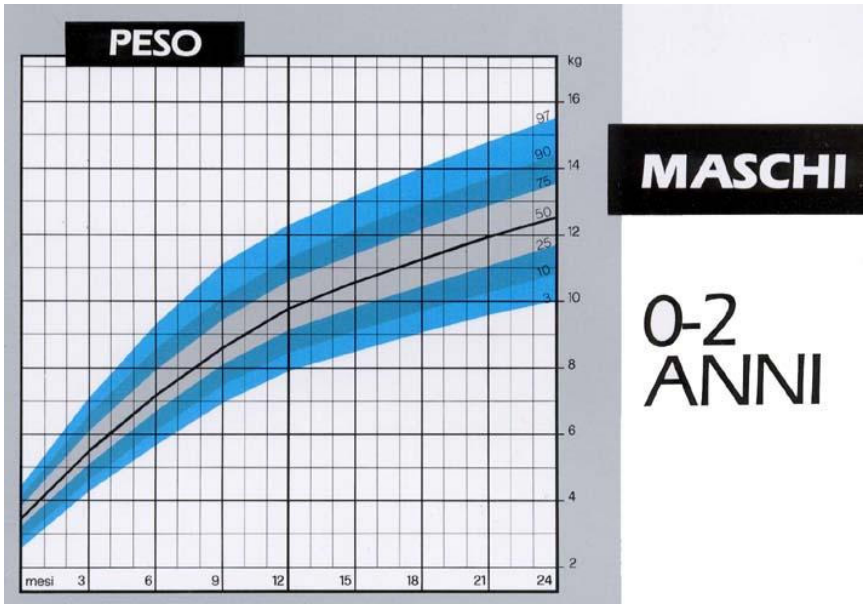
Media = 97.44

Serie:

1 1 1 2 96 560 754 930 1000

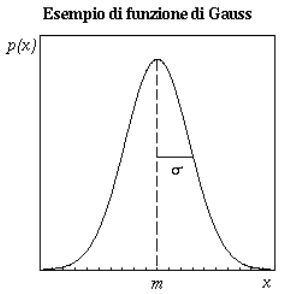
Media = 371.67

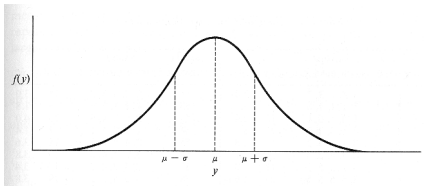
Centili



Misure di dispersione

La dispersione o variabilità è la seconda importante caratteristica di una distribuzione di dati. Essa definisce la forma più o meno raccolta della distribuzione intorno al valore centrale.





Range (campo di variazione)

$$W = x_{\max} - x_{\min}$$

- Misura puramente descrittiva e poco informativa

Es. Le altezze di 10 esemplari di una pianta sono:

10 22 33 44 46 51 67 74 79 85

W=85-10=75

Le altezze di altri 10 esemplari sono invece

10 11 11 12 13 14 15 16 20 85

W=85-10=75

Varianza di una popolazione

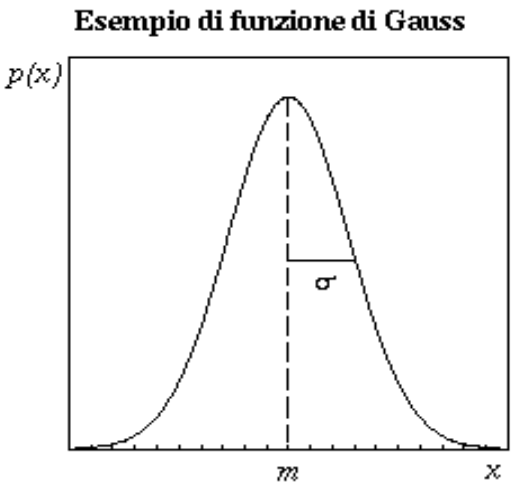
- È la media dei quadrati degli scarti tra i valori della variabile e la media.

$$V = \frac{1}{N}\sum_{i=1}^{N}\left(x_i - \overline{x}\right)^2$$

$$\sigma = \sqrt{V}$$

Si chiama deviazione standard o scarto quadratico medio

Gaussiana



Alcune formule

- Con la distribuzione

$$V = \frac{1}{\sum_{j=1}^{m} f_j} \sum_{j=1}^{m} f_j \cdot \left(\mathrm{x}_j - \overline{\mathrm{X}}\right)^2$$

- Teorema di König

$$V = \frac{1}{\sum_{j=1}^{m} f_j} \sum_{j=1}^{m} f_j \cdot \mathrm{x}_j^2 - \overline{\mathrm{X}}^2$$

Varianza campionaria

$$s^2 = \frac{1}{N-1} \sum_{i=1}^{N} \left(x_i - \overline{\mathrm{X}}\right)^2$$